ICC 2017
WASHINGTON
JULY 02-07, 2017

Multi-Scale Effects and Sensitivities in Built-up Land Data Accuracy Assessments

Johannes H. Uhl, Stefan Leyk

University of Colorado, Department of Geography, Boulder, Colorado, U.S.A.
Johannes.Uhl@colorado.edu, Stefan.Leyk@colorado.edu

Abstract: In typical accuracy assessments of cartographic products and remote sensing-based spatial data, uncertainty is often quantified using overall accuracy measures. Even though stratified, localized, and multi-scale approaches for such accuracy assessments have been proposed in the past, little research has been done regarding the sensitivity and robustness of commonly used agreement measures derived from confusion matrices at different scales using appropriate reference data covering large spatial extents. In this contribution, we explore the behavior of selected accuracy measures across the space-scale domain assessing the agreement between built-up area labels from the Global Human Settlement Layer and a reference database created by integrating publicly available cadastral and building footprint data. Administrative and census enumeration units at different levels of spatial granularity are used to generate large numbers of spatially constrained confusion matrices. Spatial and non-spatial visualization methods are employed to identify multi-scale effects and shed light on questions regarding the appropriate scale at which such data should be used.

Keywords: Accuracy assessment, multi-scale analysis, built-up land cover, spatially constrained confusion matrix.

1. Introduction

Uncertainty in geospatial data such as topographic maps or remote-sensing derived land cover data is often quanti-fied by statistical measures obtained through accuracy assessments that are based on map comparison techniques. In such assessments the examined data are compared to an independently compiled reference dataset of presumed higher accuracy. Common map comparison techniques in Cartography and GIScience include confusion matrices to derive accuracy metrics that quantify the agreement between the test data and reference data within the study area (Fielding and Bell 1997). Commonly used metrics are the Kappa index of agreement (Cohen 1960), User's and Producer's Accuracy (Story and Congalton 1986), percentage of correctly classified (PCC; Michie et al. 1994), or Normalized Mutual Information (NMI; Forbes 1995). Reference data can be generated from in-situ field measurements or from independently compiled cartographic or remote sensing data products of finer granularity (FGDC 1998). Often, such reference data are available only for few locales, and are limited to relatively small spatial extents, which requires the use of appropriate sampling techniques (Congalton and Green 1999, Stehman and Foody 2009).

In a typical accuracy assessment, an individual accuracy measure is computed for the whole study area ignoring spatial (or spatio-temporal) variation of the level of agreement between the two data sources. Hence, these aggregated statistics might misrepresent the inherent uncertainty and its spatial structure. In addition to that, it is well-known that some of these often used accuracy metrics are sensitive to the overall sample size (i.e., to the extent of the study area or geographic scale) and to the proportional size of individual classes (see Wickham et al. 2010). Accuracy metrics can be severely biased if the sample sizes across the different labels (e.g., land cover classes) vary considerably (Rosenfield and Melley 1980). To reduce these effects, different approaches have been proposed including stratified sampling (e.g., Congalton 1991), spatially constrained (localized) confusion matrices (Leyk and Zimmermann 2004, Foody 2005, Stehman and Wickham 2011), and statistical frameworks for map comparison at multiple

scales (Pontius 2002, Pontius and Suedmeyer 2004, and Pontius and Cheuk 2006) which allow to assess the spatial variation of the accuracy in different strata and at different geographic scales.

However, the effects of scale dependency and sensitivity on the outcome of accuracy assessments have not been examined using appropriate reference data covering large spatial extents. Hence, there is limited knowledge of the link between the geographic scale and fitness for use of the data for a specific purpose. For example, most land cover data are not to be used beyond the landscape scale but a specific criterion about the appropriate scale is rarely given. This is unfortunate because it limits the use of such data for local studies or provokes the use without accounting for inherent uncertainty in the data. The present study is an attempt to examine the scale sensitivity of accuracy measures that quantify uncertainty at different geographic scales.

A reference database for multi-temporal built-up areas in the U.S. is used that has been created by integrating publicly available cadastral, tax assessment and building footprint data and allows for accuracy assessment at large extents and across different time periods (see Uhl et al. 2016a, 2016b). In this contribution, we conduct an exemplary accuracy assessment of built-up area derived from the Global Human Settlement Layer (GHSL; Pesaresi et al. 2015) at different geographic aggregation levels derived from administrative boundaries and U.S. census enumeration units.

2. Method

2.1. Data and study area

Open data policy makes cadastral data, tax assessment data and building footprint data available for several regions in the U.S. Large amounts of parcel data including built year information and building footprint data were collected and an integrated data product was built that delineates built-up areas at fine spatial (building footprints) and temporal (annual) resolution. This integration process spatially refines parcel geometries to the extent of the building outlines. This valuable data source will be used to create unique snapshots of built-up land at any point in time which can be employed as large-scale reference surfaces for multi-temporal accuracy assessments of developed or built-up land classes in various land cover data products including the GHSL. Plausibility of this integrated data product is assessed by cross-comparing building and parcel information and excluding discrepant areas from the analysis, which increases the reliability of the reference data. The study area used in this work is the state of Massachusetts (U.S.) where the described reference data is available.

The test data consist of built-up areas in 2014 according to the GHSL Landsat edition, a recently released global data product that estimates the presence and distribution of human settlement on the planet at high spatial resolution (38m) and for different points in time (1975, 1990, 2000, and 2014), based on historical Landsat imagery and symbolic machine learning approaches (Pesaresi et al. 2016). Figure 1a shows the built-up labels derived from GHSL and Figure 1b shows the corresponding reference data for the whole state.

Fig. 1. (a) Built-up labels derived from GHSL, and (b) built-up labels derived from integrated cadastral and building footprint data for the state of Massachusetts in 2014 at a spatial resolution of 38m, excluding areas where reference data is not reliable.

In order to examine scale effects of measures of accuracy, we conduct accuracy assessments of built-up land in 2014 at multiple aggregation levels within the state of Massachusetts. The aggregation levels are derived from administrative boundaries (i.e., county and township boundaries; MassGIS 2016) and U.S. census enumeration units, (i.e., census tracts, block groups and blocks; U.S. Census Bureau 2016). Census tracts generally have a population size between 1,200 and 8,000 people, block groups contain between 600 and 3,000 people and census blocks represent single city blocks in urban areas, and may encompass large areas in rural regions (U.S. Census Bureau 2017).

2.2. Multi-scale accuracy assessment

Confusion matrices are built through comparison of the GHSL and reference data of the year 2014 to compute different accuracy metrics at various data-derived levels including state and county level (Figure 2a), township level (Figure 2b), and three U.S. census 2010 enumeration unit levels (i.e., tracts, block groups and blocks, Figures 2c-e). In 2010, the state of Massachusetts contains 14 counties, 351 townships, 1,475 census tracts, 4,982 block groups, and 157,508 census blocks. Since the delineation of census enumeration boundaries is heavily influenced by the underlying spatial population distribution, it can be expected that large-scale spatio-temporal patterns of population are related to those of built-up area. Therefore, using census enumeration units is an inherently meaningful way to spatially constrain the confusion matrices for substantive evaluation of underlying scale-accuracy associations. Here, User's and Producer's Accuracy are computed to exemplify the potential of this approach and assess the spatial variation of the accuracy measures at different analytical scales.

Furthermore, cross-scale links among enumeration units at different aggregation levels are established based on spatial containment criteria (e.g., spatially joining census block centroids to their containing enumeration unit at the next, more aggregated, levels). Cross-scale links will be useful to track each accuracy measure across the scale domain from the most localized scale (census block) to the universe level of the study area (state) and thus better understand the inherent variation of the accuracy measures across scale.

These vast numbers of cross-scale trajectories make it possible to examine the sensitivity of each accuracy measure to potential biasing effects related to critical sample sizes and identify numerically robust accuracy measures.

Fig. 2. Aggregation levels for the multi-scale accuracy assessment: (a) counties, and (b) townships in Massachusetts, (c) census tracts, (d) block groups, and (e) census blocks in Suffolk County.

3. Results

The results of the multi-scale accuracy assessment are presented using two visualization techniques, multi-scale choropleth maps (Section 3.1) and cross-scale trajectory plots (Section 3.2).

3.1. Spatial variation of accuracy measures across scale

Mapping the accuracy measures at different levels illustrates the inherent spatial variability at each selected scale. Figure 3 shows User's Accuracy (UA) of the GHSL built-up labels from state to census block level. Whereas UA at the state level (Figure 3a) has a similar magnitude as the majority of counties (Figure 3b), it decreases in most entities of the subsequent finer scales (Figure 3c-f), especially in rural settings. In highly urban regions (Suffolk County with the city of Boston and surrounding counties), UA tends to increase from state to census tract level but then results in segregated distributions when using units of finer granularity. Concluding, UA generated from the state level-confusion matrix tends to increasingly underestimate UA in urban settings and to overestimate UA in rural areas with increasing granularity. The low UA in rural settings is due to a high number of false positives caused by road features detected as built-up land in GHSL.

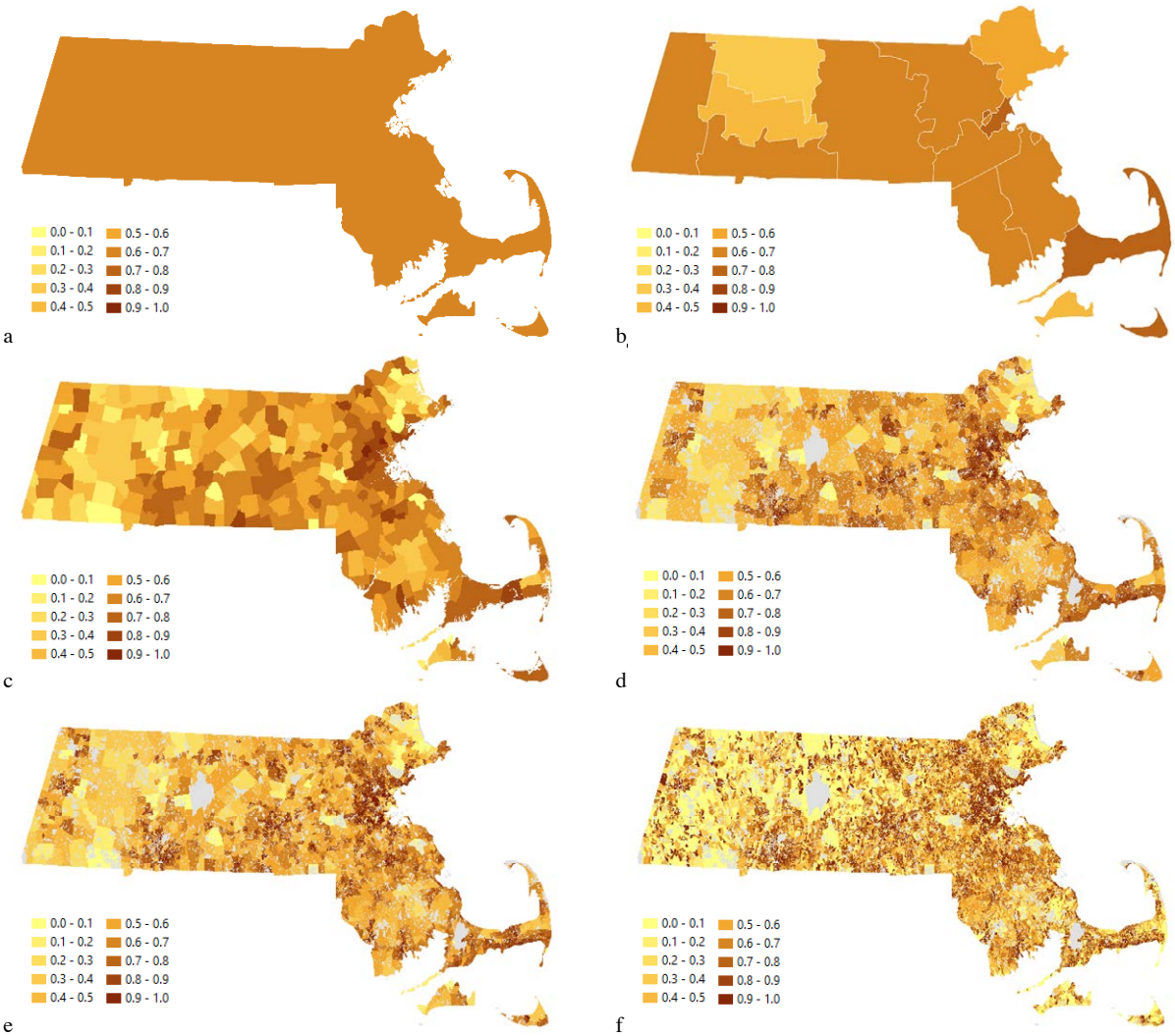

Fig. 3. Results of the multi-scale accuracy assessment, here User's Accuracy for each scale level. Gray areas are excluded from the analysis due to implausible reference data.

The multi-scale maps of Producer's Accuracy (PA) in Figure 4 show an even more distinctive pattern. In urban settings, PA increases with increasing granularity whereas PA decreases in rural areas. Similar to UA, the state-wide PA (Figure 4a) underestimates in urban areas and overestimates in rural settings when compared to finer granularity scenarios. PA at finer scales show a less segregated and thus more homogeneous pattern than was observed for UA, particularly in rural areas. Low PA values in rural areas are due to a high number of false negatives caused by the difficulty in detecting dispersed small settlements in GHSL.

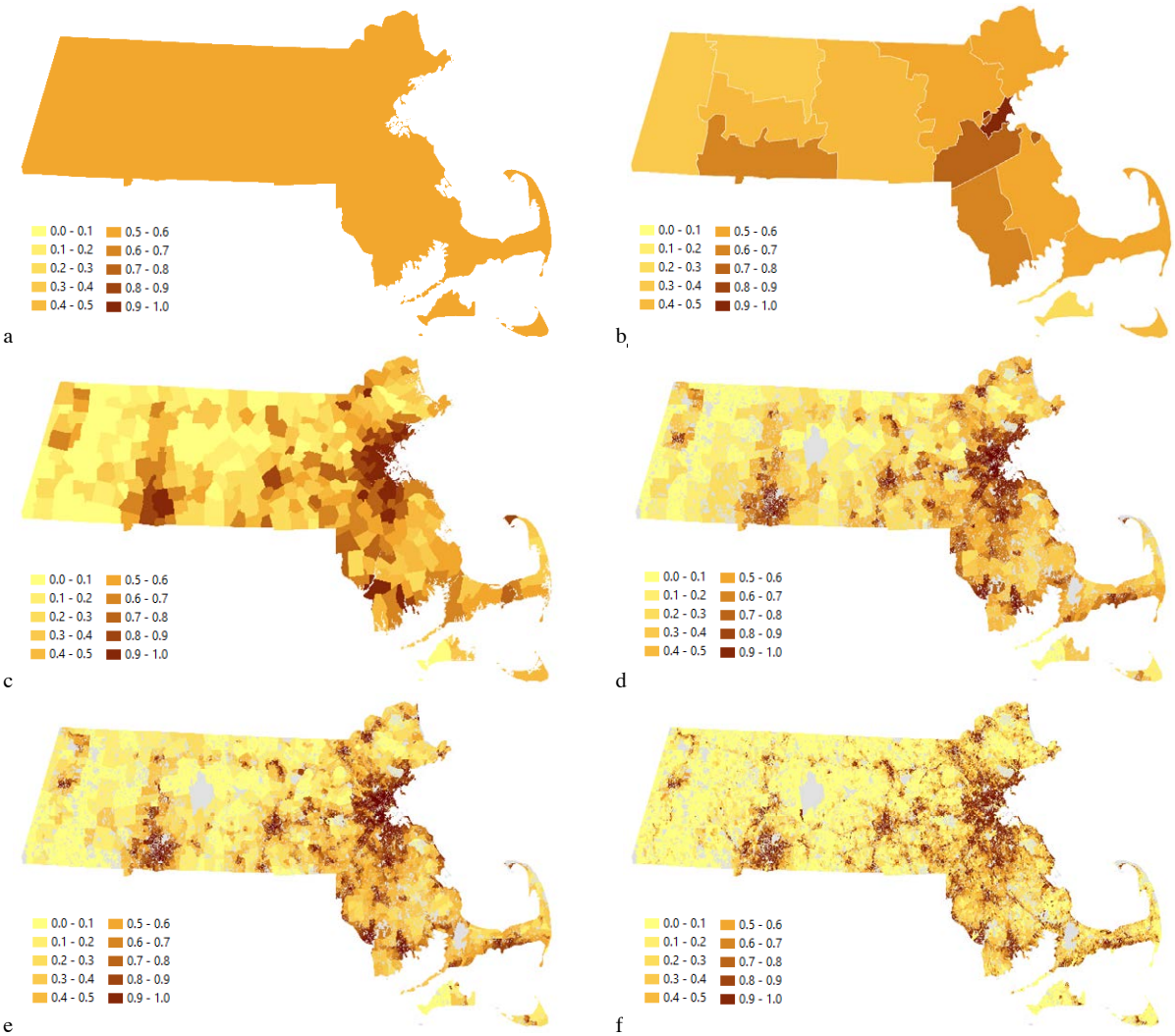

Fig. 4. Results of the multi-scale accuracy assessment, here Producer's Accuracy for each scale level. Gray areas are excluded from the analysis due to implausible reference data.

3.2. Cross-scale trajectories

Whereas multi-scale choropleth maps illustrate the spatial variability of the accuracy measures and their scale dependency, it is difficult to detect and visualize cross-scale effects. Cross-scale trajectory plots are created for UA and PA using results from selected counties. Figure 5a shows how UA varies across census blocks inside each unit at the next (coarser) level in the mostly rural Franklin County. It can be seen that county-level UA (UA_{County} = 0.360) is lower that state level UA. However, the range of UA increases steadily with increasing granularity from township to block group level. At the block level, UA values cover the full range from 0 to 1, and tend to converge to the extreme values (0 and 1) as well as to fractions of whole numbers (1/3, 1/2, 2/3, etc.) due to critically low sample sizes. This indicates that UA is not a robust accuracy measure for extremely low sample sizes.

The cross-scale trajectories for census blocks in highly urban Suffolk County (UA_{County} = 0.734) in Figure 5b show that UA stays at high magnitudes with increasing granularity whereas only in a few cases UA drops below 0.4 at the tract level. The convergence effect to extreme values and major fractions at block level is even more visible in this case.

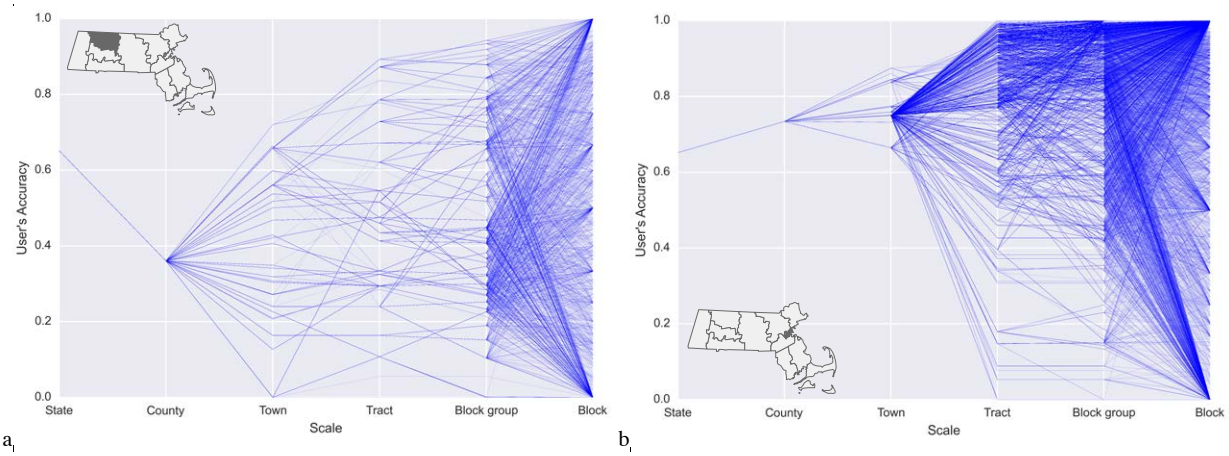

Fig. 5. Cross-scale trajectories of User's Accuracy (UA) for the census blocks contained (a) in Franklin County (mostly rural, lowest UA on county-level) and (b) in Suffolk County (highly urban, highest UA on county-level).

The cross-scale trajectories for PA are analyzed for mostly rural Dukes County ($PA_{County} = 0.235$) and for Bristol County ($PA_{County} = 0.667$) of mixed urban and rural character. In Dukes County (Figure 6a) the variation of PA increases steadily with finer granularity, whereas in Bristol county (Figure 6b) the level of variation stays relatively constant from town to block group level and then spreads to the full range at the block level.

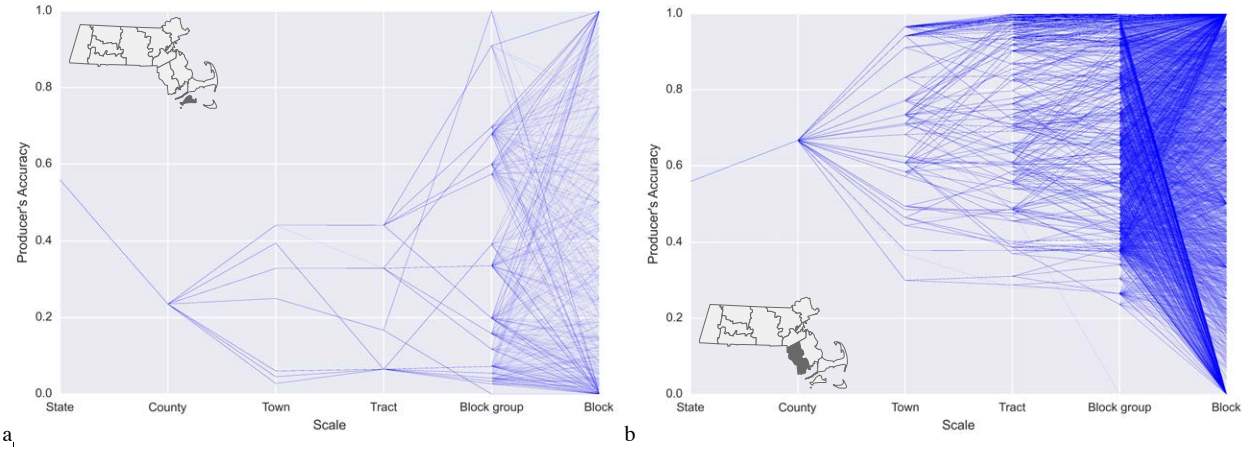

Fig. 6. Cross-scale trajectories of Producer's Accuracy (PA) for the census blocks contained (a) in Dukes county (mostly rural, lowest PA on county-level) and (b) in Bristol county (mixed urban-rural character, relatively high PA on county-level).

In all four cross-scale trajectory plots a convergence effect to the extreme values can be noted for UA and PA, as well as an apparent random dispersion of the accuracy measures at the block level. This indicates a lacking robustness of UA and PA for extremely small sample sizes related to critically small geographic extents.

4. Conclusion and Outlook

In this study, the behavior of commonly used metrics for accuracy assessment of remote-sensing derived land cover classes are analyzed across different aggregation scales using spatially constrained confusion matrices based on administrative and census-defined enumeration units of different granularity. The accuracy maps show the variability of the accuracy measures across the scale-space domain and underline the necessity of stratification or partitioning approaches for large-scale accuracy assessments and inform about meaningful geographic extents to be used

for appropriate application of the remote-sensing derived data product assessed. Cross-scale trajectories illustrate the aspatial behavior of the accuracy measures across scale and allow to infer about the robustness of the accuracy measure at a certain scale as well as along the whole scale domain. Phenomena of dispersion and convergence can be seen at the census block level, where low sample sizes begin to bias the studied accuracy measures.

These preliminary results represent a promising first step to further analysis, such as the application of statistical models to estimate the relationship between geographic extent, sample size and the variability and robustness of accuracy measures. This will aid in the exploration of alternative agreement measures with potentially higher degree of robustness against the discussed scale effects for extreme sample sizes. In the field of object-based image analysis, similar issues have been addressed by Radoux and Bogaert (2014). It should be noted that the results of accuracy assessments may also be affected by positional and thematic uncertainty in the reference data. Thematic uncertainty can be introduced by incomplete reference data or by different definitions of the map categories (e.g., land cover classes) used in reference and test data. Positional uncertainty can be introduced by inaccuracies in the registration of the reference data, different data acquisition methods or by shifts introduced through resampling or rasterization processes. Glick et al. (2016) and Uhl et al. (2016b) show how the sensitivity of accuracy metrics to positional discrepancies can be quantified using simulative approaches. Future work will analyze the interactions of such sensitivities with multi-scale effects.

Furthermore, the use of moving window techniques to assess accuracy variation at sub-block level (i.e., pixel or pixel group level) and data-driven scale levels will be investigated and combined with multi-temporal accuracy assessment approaches (Uhl et al. 2016c). In addition to that, localized accuracy assessments possess great potential for accuracy prediction by associating the obtained accuracy metrics with related ancillary demographic information (such as percentage of urban population) or landscape-related variables (such as percentage of urban area) using correlation or regression techniques. Such experiments will help to investigate the potential to predict accuracy of the data under test as a function of geographic scale and ancillary variables (Steele et al. 1998, Smith et al. 2003, Leyk and Zimmermann 2004, Zhang and May 2016) which can be applied to regions where no reference data is available.

Acknowledgements

Research reported in this publication was, in part, supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD066613. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work was also funded, in part, by the US National Science Foundation award #1416860 to the City University of New York, the Population Council, the National Center for Atmospheric Research and the University of Colorado at Boulder. Finally, Innovative Seed Grant funding from the University of Colorado to support EarthLab as well as a development grant received from its Population Center (CUPC) at the Institute of Behavioral Science, are acknowledged.

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Congalton, R.G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment* 37(1), 35-46.

Congalton, R.G., & Green, K. (1999). *Assessing the accuracy of remotely sensed data: principles and applications*. Boca Raton: Lewis Publishers.

FGDC (1998). *Geospatial positioning accuracy standards - Part 3: National standard for spatial data accuracy*. Washington, DC: Federal Geographic Data Committee.

Fielding, A.H., & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(01), 38-49.

Forbes, A.D. (1995). Classification algorithm evaluation: Five performance measures based on confusion matrices. *Journal of Clinical Monitoring and Computing*, 11, 189-206.

Foody, G.M. (2005). Local characterization of thematic classification accuracy through spatially constrained confusion matrices. *International Journal of Remote Sensing*, 26(6), 1217-1228.

Glick, H.B., Routh, D., Bettigole, C., Oliver, C. D., Seegmiller, L., Kuhn, C. (2016). Modeling the effects of horizontal positional error on classification accuracy statistics. *Photogrammetric Engineering & Remote Sensing*, 82(10), 789-802.

Leyk, S., & Zimmermann, N.E. (2004). A predictive uncertainty model for field-based survey maps using generalized linear models. *International Conference on Geographic Information Science*, 191-205.

MassGIS (2016). Office of Geographic Information, Commonwealth of Massachusetts, MassIT, http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/. Accessed 18 Aug 2016.

Michie, D., Spiegelhalter D., Taylor, C. (1994). *Machine learning, neural and statistical classification*, Ellis Horwood.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Haag, F., Halkia, M., Julea, A.M., Kemper, T., Soille, P. (2015). Global human settlement analysis for disaster risk reduction. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(7), 837-843.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, S., Julea, A., Kemper, T., Soille, P., Syrris, V. (2016). Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. *JRC Technical Report EUR 27741 EN*.

Pontius Jr., R.G. (2002). Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering and Remote Sensing*, 68(10), 1041-1050.

Pontius Jr., R.G. & Cheuk, M.L. (2006). A generalized cross□tabulation matrix to compare soft□classified maps at multiple resolutions. *International Journal of Geographical Information Science*, 20(1), 1-30.

Pontius Jr., R.G. & Suedmeyer B. (2004). Components of Agreement between categorical maps at multiple resolutions. *Remote Sensing and GIS Accuracy Assessment*, CRC Press, 233-251.

Radoux, J., & Bogaert, P. (2014). Accounting for the area of polygon sampling units for the prediction of primary accuracy assessment indices. *Remote Sensing of Environment*, 142, 9-19.

Rosenfield, G., & Melley, M. (1980). Applications of statistics to thematic mapping. *Photogrammetric Engineering and Remote Sensing*, 46, 1287-1294.

Steele, B.M., Winne, J.C., Redmond, R.L. (1998). Estimation and mapping of misclassification probabilities for thematic land cover maps. *Remote Sensing of Environment*, 66(2), 192-202.

Smith, J.H., Stehman, S.V., Wickham, J.D., Yang, L. (2003). Effects of landscape characteristics on land-cover class accuracy. *Remote Sensing of Environment*, 84(3), 342-349.

Stehman, S.V., & Foody, G.M. (2009). Accuracy assessment. *The SAGE handbook of remote sensing*, 297-309.

Stehman, S.V., & Wickham, J.D. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sensing of Environment*, 115(12), 3044-3055.

Story, M., & Congalton, R.G. (1986). Accuracy assessment - a user's perspective. *Photogrammetric Engineering and Remote Sensing*, 52(3), 397-399.

Uhl, J.H., Leyk, S., Florczyk, A.J., Pesaresi, M. (2016a). Exploring the usefulness of land parcel data for evaluating multi-temporal built-up land layers. *Proceedings of Spatial Accuracy 2016*, Montpellier, France, 95-100.

Uhl, J.H., Leyk, S., Florczyk, A.J., Pesaresi, M., Balk, D. (2016b). Exploring the potential of integrated cadastral and building data for evaluation of remote-sensing based multi-temporal built-up land layers. *Conference Proceedings AutoCarto 2016*, Albuquerque, New Mexico, USA, 212-223.

Uhl, J.H., Leyk, S., Florczyk, A.J., Pesaresi, M., Balk, D. (2016c). Assessing spatiotemporal agreement between multi-temporal built-up land layers and integrated cadastral and building data. *GIScience 2016 Short Paper Proceedings*, Montreal, Canada, 344-347.

U.S. Census Bureau (2016). Data available from https://catalog.data.gov/organization/census-gov. Accessed 18 Aug 2016.

U.S. Census Bureau (2017). *2010 Geographic Terms and Concepts*. Retrieved from https://www.census.gov/geo/reference/terms.html. Accessed 28 Feb 2017.

Wickham, J.D., Stehman, S.V., Fry, J.A., Smith, J.H., Homer, C.G. (2010). Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. *Remote Sensing of Environment*, 114(6), 1286-1296.

Zhang, J., & Mei, Y. (2016). Integrating logistic regression and geostatistics for user-oriented and uncertainty-informed accuracy characterization in remotely-sensed land cover change information. *ISPRS International Journal of Geo-Information*, 5(7), 113.